

حوار مع الاختصاصي نزار حبش حول تحدّيات الـ AI مع لغة الضّاد: الذكاء الاصطناعي بيتكلّم عربي؟

ميديا وتلفزيون | تكنولوجيا | علي عواد | الجمعة 24 أيار 2024

متى يكون للمتحدّثين بالعربية أنظمة ذكاء اصطناعي توليدية طليقة بلغتهم؟ المهمّة تبدو عسيرة بعض الشيء، لكن في حال نجاحها، فإنّ التكنولوجيا يمكن أن تلعب دوراً في توحيد لغة الضاد عبر توفير أدوات مثل المدقّق الإملائي والأدوات النحوية، ما قد يشجّع الناس على استخدام اللغة الصحيحة. وبرأي الاختصاصي نزار حبش، فإنّ ذلك سيعزّز الهوية الثقافية

بدأت أنظمة الذكاء الاصطناعي التوليدية مثل «تشات جي. بي. تي» وغيره، رسم ملامح العالم الرقمي المقبل منذ نهاية عام 2022. دخلت تلك الأدوات سريعاً في حياة الناس اليومية، لتصبح مساعداً رقمياً يسهّل ويسرّع عملهم في وظائفهم. ورغم الإثارة التي أحاطت بهذا التطور التكنولوجي، إلا أن تلك الأدوات لم تلاق المتحدثين بالعربية كمثل المتحدّثين باللغة الإنكليزية، فهي تعمل بكفاءة فقط في لغة شكسبير. ما هي الأسباب الكامنة وراء ذلك؟ ومتى يكون للمتحدثين بالعربية أنظمة ذكاء اصطناعي توليدية طليقة بالعربية؟ في محاولة لفهم ما يحدث مِن تقدّم في تدريب الذكاء الاصطناعي (AI) على اللغة العربية، أثارت «الأخبار» المسألة مع الأكاديمي الفلسطيني الأميركي نزار حبش، المتخصّص في الذكاء الاصطناعي، وتحديداً معالجة اللغات الطبيعية واللغويات الحاسوبية. عن رأيه بالوضع الحالي لجهود البحث والتطوير في أدوات الذكاء الاصطناعي المدرّبة على اللغة العربية، يقول حبش إنّه «رغم كونها متأخّرة نسبياً في مجال الذكاء الاصطناعي ومجال معالجة اللغات الطبيعية، إلا أنّ اللغة والعالم العربيّين يضعان بصمتهما كلاعب مهم في هذا المجال».



ويرجع الفضل في ذلك إلى عاملين: الأهداف والتمويل المركّز من بعض الحكومات في العالم العربي (الإمارات العربية المتحدة والسعودية)، التي تضع الكثير من الموارد في البحث والتطوير في نماذج اللغة الكبيرة LLM عبر الشركات (المحلية بالتعاون الخارجي مع اللاعبين الرئيسيين)؛ والجامعات مثل جامعة «محمد بن زايد للذكاء الاصطناعي» و«الملك عبد الله للعلوم والتقنية». ويضيف: «التحول النوعي الذي جاء مع نماذج اللغة الكبيرة LLM، سهّل بالفعل على الباحثين المؤسّسين من خارج مجتمع البرمجة اللغوية العصبية NLP، وخصوصاً الأشخاص الآتين من خلفية علوم الأنظمة والتعلّم الآلي، التركيز على هذا المجال. ما ينقصنا هو المزيد من البيانات للتدريب عليها والمزيد من مجموعات الاختبار الجديدة غير الملوّثة لتقييم السياقات العربية بشكل صحيح». يشير التلوث هنا إلى «المشكلة المتمثّلة في وجود الكثير من التركيز على توجيه البيانات إلى بناء نماذج اللغة الكبيرة، وبعض مجموعات الاختبار في التدريب، ما يؤدي إلى تحسين الأداء بشكل مصطنع (بلغة تبسيطية، يمكن تشبيه الأمر بإعطاء أجوبة الامتحان للطالب قبل الامتحان، ما يؤدي إلى حفظها والإتيان بعلامة متفوّقة). وهذه مسألة تقييم علمي مناسب».

وفي إشارة إلى التحديات والفرص الرئيسية في تطبيق تكنولوجيا LLM على النص العربي، يقول أستاذ علوم الحاسوب إنّ اللغة العربية لا تحتوي على قدر كبير من البيانات في شكل رقمي قابل للاستخدام مقارنة باللغة الإنكليزية. ولإيضاح السياق هنا، تجدر الإشارة إلى أن أنظمة الذكاء الاصطناعي تُدرَّب على كمية هائلة من النصوص (كتب، مقالات، نصوص على منصات التواصل) ومن بين تلك البيانات الموجودة، «نواجه أربع مشكلات تقلل من جودة البيانات وقابلية تعلّمها من الآلة: اللهجات، والالتباس، والتشكل، والضجيج» (يختصرها نزار حبش بـ DAMN).

في ما يتعلّق باللهجات، يقول الباحث إنّ «هناك عدداً من اللهجات العربية في شكل نصوص اليوم متوافرة على منصات التواصل الاجتماعي، والمدوّنات، التي تختلف في الأسلوب وجودة المحتوى عن اللغة العربية الفصحى التي حُرّرت بعناية.

أما الغموض، فمردّه إلى كتابة النصوص العربية من دون علامات التشكيل، ما يضيف عدداً من القراءات المحتملة (12 نطقاً/ تشكيلات/قراءات) وبمعدل 2.7 معنى أساسياً (مفردات) في المتوسط لكل كلمة عربية نموذجية، وأخيراً فالضجيج هو نتاج الأخطاء الإملائية في اللغة العربية بنسبة كبيرة (الهمزات والتاء المربوطة). وجدنا أن 30% من الكلمات على وسائل التواصل الاجتماعي التي تنوي الكتابة باللغة العربية الفصحى تحوي أخطاء. تتفاعل العناصر الأربعة لأنّ اللهجات ليس لديها قواعد إملائية رسمية، ويمكن أن تكون أكثر تعقيداً في بعض الأحيان من المعيار وحتى غامضة عبر متغيّرات اللهجات والمعايير. كل هذا يجعل من تعلّم نماذج LLM أكثر صعوبة. أما من ناحية الفرصة، فإنّ حل بعض هذه المشكلات يُعد أمراً جيداً للذكاء الاصطناعي باللغة العربية بشكل عام وسيكون له مفاعيل للغات الأخرى».

وعن التأثير المحتمل لتطوير أنظمة AI تتقن العربية على مختلف القطاعات مثل التعليم والإعلام والتكنولوجيا في العالم الناطق باللغة العربية، يقول نزار حبش إنّ كل تلك القطاعات ستستفيد بالتأكيد، لكن الثقافة العربية ستستفيد بطريقة فريدة من نوعها، «ففي حين أنّ الناس يستخدمون عادة لهجاتهم المحلية بدلاً من اللغة العربية الفصحى الرسمية في الحياة اليومية، إلا أن التكنولوجيا يمكن أن تلعب دوراً في تعزيز توحيدها عبر توفير أدوات مثل المدقق الإملائي والأدوات النحوية. وهذا يمكن أن يشجع الناس على استخدام اللغة الصحيحة ووضع معايير غير رسمية تدريجاً. وفي نهاية المطاف، يُنظر إلى هذا التوحيد بشكل إيجابي بالنسبة إلى العالم العربي لأنه يعزّز الثقة اللغوية، والاعتزاز بالهوية الثقافية، والوعي الذاتي مع تقدم المنطقة نحو المستقبل».